Measuring Economic Growth from Outer Space: A Comment

Marcus Berliant and Adam Weiss

Washington University in St. Louis, The Urban Institute

20 April 2017

Measuring Economic Growth from Outer Space: A Comment*

Marcus Berliant† and Adam Weiss‡

April 2017

Abstract

We examine econometric and elementary economic theory issues arising from the model specification in Henderson, Storeygard and Weil (2012), that uses night light data to proxy for missing or unreliable GDP growth data. An alternative approach based on the expenditure function is outlined. It can accommodate prices as well as quantity information from other commodity markets.

JEL numbers: D11, D61, O47, O57 Keywords: GDP, Night light data, Omitted variable, Expenditure function, Spatial autocorrelation

*The authors thank Eckhardt Bode, J.V. Henderson, T. Holmes, R. Jedwab, R. Manuelli, D.P. McMillen and D.N. Weil for helpful comments and Andrés Hincapie Norena for superb research assistance, implicating none of them for the contents of this paper.

†Department of Economics, Washington University, Campus Box 1208, 1 Brookings Drive, St. Louis, MO 63130-4899 USA. Phone: (314) 935-8486, Fax: (314) 935-4156, e-mail: berliant@wustl.edu

‡The Urban Institute. e-mail: weiss.l.adam@gmail.com

1 Introduction

Henderson et al (2012) employ night light data to augment more standard measures of GDP growth, particularly for countries where data is unreliable or missing, or for subnational units. Here we examine some issues that arise with this approach. Our comments come both from the purely empirical point of view as well as from elementary consumer theory. Elementary demand theory strongly suggests that the price of electricity should play an important role as a right hand side variable either in regressions of lights on GDP or the predictive regression of GDP on lights. Let us be more specific.

In the first regression of lights on GDP, essentially a demand function where GDP plays the role of income, the omitted price variable can cause spatial autocorrelation of the errors (if electricity price is correlated across space) and thus bias in the coefficient on GDP. In the predictive regression of GDP on lights, light usage can be negatively correlated with the price of lights, in the error term, thus leading to bias in the coefficient on lights.

We take two approaches to address this problem.

The first, our preferred and new approach, derives the regression of GDP on lights using the expenditure function and inserts the omitted variable, the price of electricity, as a right hand side variable. This eliminates the omitted variable bias. There are two subsidiary issues that we can address with this approach. They both relate to the econometric theory. Call the estimate of the coefficient from the regression of night lights on GDP β. First, discounting the omitted variable bias just mentioned, as is recognized in Henderson et al (2012), there is bias in using β to form the estimate of $\frac{1}{\beta}$; see equation (9) below. This is accounted for in the actual estimates of $\frac{1}{\beta}$, but the bias correction is not made before minimizing variance to form an estimate of GDP. We do not have this problem. Second, in Henderson et al (2012) fixed effects are used in the panel regressions (where the data units are country-year pairs), but of course not in the long term difference regression (where the data units are countries). The fixed effects are not inserted into the econometric theory for the panel data, so their use is not justified. We justify use of fixed effects in the panel regression.

Second, as an alternative we consider a spatial econometric approach. Here the omitted price variable is proxied using the spatially lagged independent variable, night lights. If trade across countries is also important, then spatially lagged GDP should also be included in the regressions because higher income in one country implies higher demand for an adjacent country's products and

thus higher income for that neighbor.

Our empirical findings are as follows. First, without inserting the price variable in the regression of GDP on lights, there is strong evidence of spatial autocorrelation in the errors. Second, when price is inserted into the regression, it is an important predictor of GDP and its elasticity is slightly lower than the elasticity for lights, but it is estimated more precisely than the lights elasticity.

Before proceeding, we wish to emphasize that *although supply and demand are clearly relevant to our work, they do not easily incorporate GDP into the analysis as a dependent variable.* That is why we use the expenditure function approach. *We are not estimating equilibrium quantities or prices as dependent variables*, though they are obviously related to our exercise. Discussion of endogeneity and causality is deferred to the end of Section 3, as we first need to develop our framework.

We begin with a recapitulation and analysis of the logic in Henderson et al (2012) in Section 2, turning later to the empirics in Sections 3 and 4. Section 3 considers our preferred approach, namely the expenditure function approach, whereas Section 4 presents an alternative, the spatial econometric approach. We end with the implications of our analysis in Section 5.

2 Analysis

To begin, we reiterate the relevant notation from the paper that is our focus. Let y be the growth in true, real GDP, and let x be the growth in night lights. Let z denote measured growth in GDP. The subscript j denotes a country. Equation (1) of Henderson et al (2012) gives the specification of measurement error in the data:

$$z_j = y_j + \varepsilon_{z,j} \tag{1}$$

where $\varepsilon_{z,j}$ is measurement error. Equation (2) of Henderson et al (2012) is the statement of a basic relationship:

$$x_j = \beta y_j + \varepsilon_{x,j} \tag{2}$$

Here, $\varepsilon_{x,j}$ is the error term, whereas β is the elasticity of (growth in) lights with respect to (growth in) income. This equation is interpreted by Henderson et al (2012) as a purely statistical relationship.

In contrast with that paper, we interpret this as a statement of the time derivative of the (log) demand relationship for light. Suppose that X is light

consumption, Y is income, and P is the price of electricity (in general, we denote levels of a variable by a capital letter and growth rates by a lower case letter). Suppose further that demand takes the functional form:

$$\begin{aligned} X &= Y^{\beta} \cdot P^{\alpha} \\ \beta &> 0,\ \alpha < 0 \end{aligned} \tag{3}$$

Taking logarithms of both sides and then the derivative with respect to time, and denoting growth rates by lower case variables, we obtain:

$$x = \beta y + \alpha p \tag{4}$$

There is an omitted variable relative to equation (2), namely the percent change in the price of electricity, the same as the percent change in the price of light. Then we would like to write

$$\begin{aligned} x_j &= \beta y_j + \alpha p_j + \varepsilon'_{x,j} \\ &= \beta y_j + \varepsilon_{x,j} \\ \text{where } \varepsilon_{x,j} &= \alpha p_j + \varepsilon'_{x,j} \end{aligned}$$

Since z denotes the measured growth of GDP, y is replaced with z from equation (1):

$$x_j = \beta z_j + \alpha p_j - \beta \varepsilon_{z,j} + \varepsilon'_{x,j} \tag{5}$$

Equation (6) is the basic relationship estimated in Henderson et al (2012, Table 2):

$$z_j = \widehat{\psi} x_j + e_j \tag{6}$$

In terms of our notation, and inserting a time index t for clarity (as panel data is used), we obtain the following expression from equation (5):

$$z_{jt} = \frac{1}{\beta} x_{jt} - \frac{\alpha}{\beta} p_{jt} + \varepsilon_{z,jt} - \frac{1}{\beta} \varepsilon'_{x,jt} \tag{7}$$

In other words, in equation (6),

$$e_{jt} = -\frac{\alpha}{\beta} p_{jt} + \varepsilon_{z,jt} - \frac{1}{\beta} \varepsilon'_{x,jt}$$

Before moving on to describe estimation, there is a very important insight that relies on this specific functional form, and is implicit in Henderson et al (2012). Viewing the system as deterministic, equation (4) represents growth in demand. But we could just as easily take logarithms of the level equation (3) to obtain:

$$\ln(X) = \beta \ln(Y) + \alpha \ln(P)$$

Taking derivatives with respect to time,

$$x = \beta y + \alpha p$$

Notice that α and β are the same no matter whether we use levels or growth rates, given the functional form assumption. In Henderson et al (2012), Table 2 is constructed using regressions on levels rather than growth rates; naturally, in Henderson et al (2012), α is set to 0. Of course, once one moves to estimation, the error structure and estimates differ depending on which equation, levels or growth rates, is used. We also note that in the actual estimates used to augment GDP data, long term differences of logarithms (averaged for two years at the beginning and end)[1] and thus growth rates are employed (Henderson et al 2012, Table 4, column 3).

The next step of model development in the paper is to construct a composite estimate of GDP that relies on available data and the estimate of GDP from night lights; the latter is constructed from equation (6) as $\widehat{z}_j = \widehat{\psi} x_j$. Thus, defining λ as the weight on available data, the estimate of GDP $\widehat{y}_j$ is formed in equation (5) of Henderson et al (2012) as follows:

$$\widehat{y}_j = \lambda z_j + (1-\lambda)\widehat{z}_j$$

Then λ is chosen so as to minimize the variance of $\widehat{y} - y$. An issue with this methodology is that $\widehat{y}$ is a biased estimator of y. To see this, using equation (2) compute

$$E(\widehat{y} - y) = (1-\lambda)[\widehat{\psi}\beta - 1]E(y) \tag{8}$$

A key equation for the analysis (equation (4) of Henderson et al (2012)) is the calculation:

$$\text{plim}\widehat{\psi} = \frac{1}{\beta}\left(\frac{\beta^2\sigma_y^2}{\beta^2\sigma_y^2 + \sigma_x^2}\right) \tag{9}$$

where σ_y^2 is the variance of true income growth and σ_x^2 is the variance of ε_x, defined in equation (2). From this equation, we can see that the bias is usually not zero. In general, one would probably want to construct an estimator that corrects for this bias in $\widehat{\psi}$ and thus in the constructed estimate $\widehat{z}$ *before* trying to minimize variance to obtain λ. This can be accomplished, for example, by using sample moments. To pursue this a little further, notice that

$$\text{plim}\widehat{\psi}\beta = \frac{\beta^2\sigma_y^2}{\beta^2\sigma_y^2 + \sigma_x^2}$$

[1] Note that the average of the logarithms is used rather than the logarithm of the average. We have remained consistent with that in our regressions.

From equation (9b,c) of Henderson et al (2012), the numerator is the sample moment $cov(x, z)$, whereas the denominator is the sample moment $var(x)$. A bias correction can likely be based on substituting this into equation (8), using the sample mean of z in place of $E(y)$, and by subtracting this estimated bias from $\widehat{z}_j$ (or from $(1-\lambda)\widehat{z}_j$). The approach likely leads to intractability of the expression for the variance of $\widehat{y}-y$, so we shall not pursue it further. Hence, we shall take a different approach.

Before leaving this topic, let us be clear. The coefficient $\widehat{\psi}$ is a fine but biased estimator of $\frac{1}{\beta}$. That is not our concern. Rather, we are concerned about how λ is estimated. It is calculated by minimizing the variance of $\widehat{y}-y$. The issue we have is that λ is not chosen by minimizing the variance *subject to the constraint that the resulting* $\widehat{y}$ *is an unbiased estimator*, namely $E(\widehat{y}-y)=0$. Thus, in general, the λ that is chosen to minimize variance will imply $E(\widehat{y}-y)\neq 0$.

Finally, notice that expression (9) implies that $\widehat{\psi}$ is an OLS estimator without fixed effects (or even a constant). Equation (9) is used to calculate, for example, the optimal value of λ. If fixed effects were included, then expression (9) would change. This is fine for the long difference regressions that use countries as units (Table 4 of Henderson et al (2012)) where the long differences (column 3) have no fixed effects, but does not justify the inclusion of fixed effects in the panel regressions (Table 2 of Henderson et al (2012)). Of course, the latter regressions are used more as motivation rather than for augmenting GDP data. Our new approach can accommodate fixed effects as well as a constant in the panel regressions.

Next we complete the description of the analysis in Henderson et al (2012). The main regression applied for estimates of GDP is equation (6) using long differences in variables rather than levels. This can be found in column 3 of Table 4 in that paper. The sample moments and a signal to noise ratio yield values for the variables of interest, including λ and β. This is applied to situations where there is bad data, but data nonetheless, on GDP. For regions where there is no data, the following technique is used. The regression with long differences, no fixed effects and a biased estimate $\widehat{\psi}$ is applied to lights data to form an estimate of GDP.

To summarize, there are three issues derived from this analysis that we intend to deal with: the constraint $E(\widehat{y}-y)=0$ should be imposed when minimizing the variance of $\widehat{y}-y$ to obtain the optimal value of λ, the omission of fixed effects in the econometric theory even though they are present in

the panel regressions, and most importantly the omission of a relevant price variable. Rather than deal with each of these piecemeal, we introduce an entirely different approach.

3 The Expenditure Function Approach

We shall see in the next section that there is evidence of spatial autocorrelation in the residuals of regression (6), indicating a problem. Evidently, we should insert the price of electricity into the regression, as suggested at the end of the last section.

To address the three issues listed at the end of the last section, we use a completely different approach, based in public finance. Let the quantity of other goods consumed be denoted by W and let the utility function $U : \mathbb{R}^2_+ \rightarrow \mathbb{R}$ be:

$$U(X, W) = \min(X, W)$$

There are several remarks to be made about this utility function. First, it is not entirely unnatural to assume that night lights are complementary to other commodities. Second, this type of utility function fulfills the assumptions used in Jerison (1984) for aggregation of individuals to a representative consumer, so it is not necessary to assume directly that an aggregate consumer exists. Third, one might view this assumption as the real, but implicit, rationale for using lights as a proxy for welfare or GDP - that night light consumption varies in proportion to welfare. From the derivations that follow, clearly this assumption about the form of utility can be generalized, but at the expense of a more complicated regression and further data requirements. We shall discuss this briefly in the conclusions.

Other goods are chosen as numéraire. The expenditure function is defined as:

$$E(P, u) = \min_{X,W} P \cdot X + W \text{ subject to } \min(X, W) = u$$

The solution to this problem sets $X = W = u$. Hence

$$E(P, u) = E(P, X) = P \cdot X + X = (1 + P) \cdot X$$

Using the identity that expenditure is income from this macro viewpoint,

$$Y = E = (1 + P) \cdot X$$

or

$$\ln(Y) = \ln(1 + P) + \ln(X) \tag{10}$$

The final form of the regression we use as an analog to the panel regressions in Henderson et al (2012, Table 2) is:

$$\ln(Z) = \text{constant} + \alpha \cdot \ln(1+P) + \beta \cdot \ln(X) + \text{panel fixed effects} + \text{error} \quad (11)$$

A few remarks are in order, particularly about the appearance of coefficients α and β. With respect to β, if X is light consumption, X^β is the electricity used to generate the light. The idea is that the technology used for generating night light from electricity is not necessarily constant returns at one to one. In fact, we would expect some heterogeneity in the technology across both time and location, and light production from electricity might be subject to either increasing or decreasing returns. Similarly, if P is the average cost of electricity, $(1 + P)^\alpha$ is the marginal cost of electricity for lights. Since electricity is often priced in a non-proportional fashion, and other uses such as running refrigerators might consume the infra-marginal units, this represents an effort to translate average to marginal cost. The theory gives us $(1+P)$ in place of P. Intuitively, α represents the elasticity of GDP with respect to the *relative* price of electricity, $\frac{\Delta P}{1+P}$, in other words translating the average relative price to the marginal relative price.

One final, but important, remark should be made about the elementary theory. Of course, with this utility function, the uncompensated price elasticity of demand for electricity should generally be negative. However, the point of our exercise is not to estimate this elasticity, but rather to estimate the expenditure function. Since X is a proxy for utility level, one would expect β to be positive. But given a utility level, the higher the price of electricity, the higher is the expenditure in terms of numéraire to attain the given utility level. In other words, one would expect α to be positive, because we are using compensated demand in conjunction with the expenditure function; the derivative of the expenditure function with respect to price is Hicksian or compensated demand, which is generally positive. Of course, as economic growth occurs, we expect to see shifts in the supply curve for electricity (or in the case of a monopolistic or regulated industry where the supply curve is irrelevant, changes in price). But this simply helps us trace out the expenditure function.

The application to contexts with missing GDP data is straightforward, given our approach. For the application that supplements bad GDP data with night light estimates, we make an assumption analogous to one in Henderson et al (2012), namely that the errors in equations (1) and (11) are uncorrelated. Then choosing λ to minimize the variance of $\widehat{y} - y$ amounts to choosing the estimate, either from regression (1) or regression (11), where the residual variance

is lowest. We can obtain an estimate of the residual variance for regression (1) by using sample moments as described in Henderson et al (2012) equations (9a,c).

The derivative of regression (11) with respect to time, interpreted as long differences, is:

$$\Delta \ln(Z) = \text{constant} + \alpha \cdot \Delta \ln(1+P) + \beta \cdot \Delta \ln(X) + \text{error}$$

or in terms of growth rates:

$$z = \text{constant} + \alpha \cdot (1+p) + \beta \cdot x + \text{error}$$

where $(1+p)$ denotes $\frac{\Delta P}{1+P}$. We use the first formulation in the regressions.

Our data on prices comes from the International Energy Agency (IEA), data.iea.org. Additional prices can be found at www.eia.gov/countries/prices/electricity_households.cfm, and sources cited therein. These sources can be useful for updating the regressions or providing data on countries not covered by the IEA data. The data from the IEA is denominated in US dollars per megawatthour. It is an unbalanced panel, as the availability of data across countries and years is not systematic. We use household prices rather than industrial prices, consistent with our theory as well as the idea that households are perhaps more likely to use light at night than firms. The IEA web site also has data on industrial prices as well as prices based on PPP.

Table 1[2]

Dependent Variable - Real GDP

	ln(GDP)	ln(GDP)	ln(GDP)
	(1)	(2)	(3)
	OLS	OLS restricted sample	OLS
ln(lights/area)	0.277***	0.142	0.157*
Standard error	[0.031]	[0.096]	[0.083]
ln(price+1)			0.086**
Standard error			[0.033]
Observations	3,015	493	493
(Within country) R^2	0.769	0.865	0.876

[2]The Stata code, price data and spatial weight matrix used to generate the tables are available from the authors upon request.

Notes: All regressions include a constant and space and time fixed effects. Standard errors are robust, clustered by country.

$***$ Significant at the 1% level.

$**$ Significant at the 5% level.

$*$ Significant at the 10% level.

Before discussing these results in detail, a remark is in order. Real GDP or real growth in GDP is used as the dependent variable. We are using nominal prices as an independent variable. This raises the obvious issue of inflation. Since the price of electricity is given in dollars, inflation of local currency relative to the dollar is accounted for in exchange rates. Inflation in the dollar over time is accounted for in the time fixed effect in these log-log regressions. For the long difference regressions, this is picked up by the constant.

Please refer to Table 1 for the results from our regressions. The first column is a replication of Henderson et al (2012) Table 2 column (1). This is the basic regression of ln(GDP) on ln(lights/area). Column (2) of Table 1 gives the same regression but for the unbalanced panel restricted to time-country pairs for which we have price data. Given the smaller data set, it is unsurprising that the coefficient on lights becomes less precisely estimated. It is somewhat surprising that the within country R^2 rises; the time fixed effects are picking up more of the variation. In column (3) of Table 1, we run the same regression as in column (2) but with prices inserted as an independent variable. Both the lights and price coefficients are significant (with the smaller sample size). The price elasticity is positive, with about half the magnitude of the lights elasticity. As one would expect with the addition of a new independent variable, R^2 rises.

Next we move to the long differenced form of the regressions, corresponding to Table 4 of Henderson et al (2012).

Table 2

Dependent Variable - Growth in Real GDP

	Δln(GDP)	Δln(GDP)	Δln(GDP)
	(1)	(2)	(3)
	OLS	OLS restricted sample	OLS
Δln(lights/area)	0.327***	0.162	0.172*
Standard error	[0.046]	[0.115]	[0.096]
Δln(price+1)			0.122**
Standard error			[0.049]
Observations	113	27	27
R^2	0.300	0.060	0.180

Notes: Robust standard errors in brackets. Long differences are formed by averaging the first and last two years of data (the first two years are 1992/1993, the last two years are 2005/2006).

∗ ∗ ∗ Significant at the 1% level.

∗∗ Significant at the 5% level.

∗ Significant at the 10% level.

The first column in our Table 2 replicates the long difference regression in Table 4 of Henderson et al (2012), namely column (3). This is the estimate used in applications. The observations are low and middle income countries. Column (2) of our table shows the same regression with a sample restricted to those countries with price data in the first two and last two years. It is important to remark here that due to sample size considerations, we are not restricting our sample to low and middle income countries. (Henderson et al (2012, p. 1016) remark that without this sample restriction, the overall estimate of the coefficient on lights is .321, with a standard error of 0.042.) Of course, it would be possible to expand our sample by not restricting it to countries with data for all 4 years, but the resulting regression would not be analogous to the regression in Henderson et al (2012).

In column (2), the R^2 falls and the coefficient on lights becomes insignificant. Finally, column (3) contains the regression with price and the restricted sample. Both lights and price have significant coefficients, with the lights elasticity a bit higher but the price elasticity estimated more precisely. The R^2 jumps up for this smaller sample with the addition of the price variable.

We suggest this last regression (column (3)) as a replacement for our column (1), used in Henderson et al (2012).[3]

In fact, our empirical approach understates our concerns about the omission of the price variable in the regressions.[4] We are worried that lights are not a perfect proxy for GDP because countries vary in electricity price as well as night lights, and price affects demand as well as expenditure. Naturally, a change in price generates both income and substitution effects. In our empirical approach, through the choice of utility function, we have shut down substitution effects and considered only income effects. To the extent that there are substitution effects, our concern about omission of the price variable is still valid, but we are solving only part of the problem since we are leaving the issue of substitution effects aside. If expenditure on electricity is only a small part of the budget for consumers, then income effects should be small relative to substitution effects. This could all be resolved by estimating an expenditure function implied by a utility function allowing substitution effects, such as Cobb-Douglas, but that would require more data, as we shall discuss in the conclusions.

By now, the reader is likely apoplectic because we have not yet addressed endogeneity and reverse causality in the regressions. It is quite evident that all the variables we have discussed are endogenous. That holds even for the most basic relationship, regression (2), because lights are a component of GDP. If equation (10) holds, the problem is exacerbated. As noted in Henderson et al (2012, footnote 14), instruments are hard to come by. But hints about this issue can be found by comparing the short run panel regressions with the long difference regressions, as follows.

Reverse causality in our regressions would mean that rising incomes, as measured by GDP, would cause an outward shift in demand for lights, moving up along the supply curve, reflected in both more use of lights and a higher price. In the long run, it would also mean a shift outward of the supply curve, reflected in lower prices and more lights. In the short run, we expect the supply of electricity to be relatively inelastic compared with the long run. Thus, if there is reverse causality in our regressions, one would expect a short run price rise and not much change in quantity, as the outward shift of demand is reflected mostly in a higher price as opposed to an increased quantity. In the long run, one expects a more elastic supply curve (as well as an outward

[3]Notice that if the price of electricity varies over time, this will not be picked up by country fixed effects.

[4]This remark originates with Tom Holmes, though he should not be blamed for it.

shift in the supply curve) deriving from capacity increases. Thus, in the long run, one expects higher quantity responsiveness and lower price responsiveness compared to the short run. In fact, the positive coefficient on price is smaller for the panel regressions than for the long difference regressions. That is opposite the prediction one would have when reverse causality is present.[5] This indicates that reverse causality is limited.

4 The Spatial Econometric Approach

One drawback of the approach outlined in the previous section is that price data is not universally available. We shall return to this issue in the conclusions. But this provides motivation for an alternative approach that does not rely on price data. It has its own drawbacks, that we shall outline, but first we must be more specific.

We wish to raise two spatial econometric issues with regression (6). For simplicity, we shall discuss level regressions here rather than growth regressions.[6]

First, although the price of electricity is not observed, it might be correlated between, for example, neighboring countries i and j. If P_{it} and P_{jt} are correlated, then there is a classical spatial autocorrelation/omitted variable problem. This results, for example, in a biased estimate of the key parameter $\widehat{\psi}$ *provided that* X_j and P_j are correlated.[7] Of course, given a demand equation, it is quite natural that use of lights be (negatively) correlated with the price of electricity. Without price data to insert into the regressions, if light use is negatively correlated with the price of electricity in a location, and electricity price in one location is positively correlated with the electricity price in neighboring locations, then we can employ spatially lagged light use to proxy for these correlations, particularly for the omitted variable: the price of electricity.

Second, it seems clear to us that due to trade between countries close in distance, real GDP at a given time could be correlated across space. In

[5] In fact, the responsiveness of lights to GDP is slightly higher in the long difference regressions than in the panel regressions, indicating the possibility of reverse causality, but the difference is rather small.

[6] One could just as easily frame the discussion adding "growth in" before each occurrence of a variable such as GDP, night lights or price.

[7] As pointed out in Henderson et al (2012), there is another, independent reason $\widehat{\psi}$ might be biased. See equation (9).

particular, higher income in one country can lead to higher demand for an adjacent country's products, thus raising income in the adjacent country. In other words, Y_{it} and Y_{jt} are correlated. This can also lead to biased estimates due to the omission of spatially lagged (and weighted) endogenous left hand side variables from the right hand side of the regression. Evidently, this issue does not arise in regressions where lights appear on the left hand side of the regression, but it does arise when GDP is put on the left hand side whereas night lights are moved to the right hand side. This second issue is covered by inserting spatially lagged dependent variables on the right hand side of the regression.[8]

Formally, combining the preceding two paragraphs with equation (7) where lagged lights proxy for price amounts to:

$$\ln(Z_{jt}) = \widehat{\psi} \ln(X_{jt}) + e'_{jt}$$

where

$$e'_{jt} = \mu \ln(Z_{it}) + \frac{\gamma}{\beta} \ln(X_{it}) + \varepsilon_{z,jt} - \frac{1}{\beta} \varepsilon'_{x,jt}$$

and countries i and j are neighbors.

Our goal is to address these issues, beginning with the basic regressions run in the paper, to see what effect this has on the use of light data to predict GDP. Our first focus is on reconstructing column 1 of Table 2 in Henderson et al (2012). Then we shall draw the implications for the analysis.

What we have is a model that incorporates both the omitted price variable indirectly using the spatially lagged independent variable night lights as a proxy, and the spatially lagged dependent variable on the right hand side. Recalling that panel data is used, we write the econometric model as:[9]

$$\ln(Z_t) = \mu W \ln(Z_t) + \ln(X_t)\psi + W \ln(X_t)\frac{\gamma}{\beta} + u_t \tag{12}$$

where W is a spatial weight matrix, Z_t and X_t are the vectors of cross sections of the respective variables at time t, and $u_t (= \varepsilon_{z,jt} - \frac{1}{\beta}\varepsilon'_{x,jt})$ is the error vector.

Of course, there are important drawbacks to this approach. One must believe that the specification of W is correct, and that the error term is no longer correlated with the new regressors.

[8] In addition, inclusion of spatially lagged GDP will pick up other unobserved spatial spillovers that cause correlation in GDP.

[9] This is sometimes called the spatial Durbin model. The taxonomy of spatial econometric models seems unsettled. To be consistent, we are using the terminology of the Stata packages we employ for estimation below, namely spglsxt and spregfext. This terminology might be confusing for some.

An important special case of (12) is when $\gamma = 0$, namely there is spatial autocorrelation only in the dependent variable. This special case is called the Spatial Autoregressive Model (SAR):

$$\ln(Z_t) = \mu W \ln(Z_t) + \ln(X_t)\psi + u_t$$

A second important special case is when $\mu = 0$, so (12) reduces to:

$$\ln(Z_t) = \ln(X_t)\psi + W \ln(X_t)\frac{\gamma}{\beta} + u_t$$

This is called the Spatial Durbin Model, or SDM.[10]

Tables 3 and 4 contain our empirical findings. The first column replicates the basic regression of Henderson et al (2012), column (1) of their Table 2. Unfortunately, we are unable to use their entire sample, as data is missing for some countries in some time periods, rendering the spatial weighting matrix that we must use to test for spatial econometric purposes difficult. Thus, we censor the countries for which data is incomplete, resulting in a smaller cross section sample size of 150. So for comparison purposes, in column (2) of our Table 3 we perform the same regression as in column (1) but for the smaller sample size. In Table 4, we report spatial test statistics for the appropriate regressions in Table 3. Throughout our application, the spatial weighting matrix that we use is simply a contiguity matrix with elements 0 and 1, where 1 is used to denote a geographic neighbor and 0 is used to denote the complement.[11] For the remaining regressions/columns, a spatial weighting matrix is necessary. In column 3 of Table 3, we run the GMM version of column 2 that also generates test statistics for model specification.[12] We find strong evidence of positive spatial autocorrelation in the error terms of this regression, as seen (for example) in Moran's I statistic. In column 4, we run the GMM correction for the misspecification, essentially equation (12), incorporating the possibility of spatial autocorrelation in the fixed effects.[13] In column 5, we run a SAR with fixed effects, whereas in column 6, we run the SDM model with fixed effects.

[10] This is called the spatial lagged X model by some.

[11] It is possible that the use of geographic distance in place of this crude measure might yield different results.

[12] Specifcally, we run the initial GMM version of spglsxt.

[13] By this we mean that a specific country can have a common affect on all its neighbors, independent of time. Operationally, we run the fully weighted GMM version of spglsxt, with the time and space fixed effects entered manually as right hand side variables.

Table 3

Dependent Variable - Real GDP

	ln(GDP)	ln(GDP)	ln(GDP)	ln(GDP)	ln(GDP)	ln(GDP)
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS restricted sample	GMM initial unweighted	GMM[14] fully weighted	SAR	SDM
ln(lights/area)	0.277***	0.267***	0.267***	0.267***	0.301***	0.369***
Standard error	[0.031]	[0.031]	[0.011]	[0.011]	[0.033]	[0.041]
ln(lagged GDP)					0.174***	
Standard error					[0.014]	
ln(lagged lights/area)						0.100***
Standard error						[0.016]
Observations	3,015	2,550	2,550	2,550	2,550	2,550
Countries	188	150	150	150	150	150
R^2	0.769	0.772	0.9994	0.9994	0.9990	0.9983

Notes: All regressions include a constant and space and time fixed effects. Standard errors are robust, clustered by country. The GMM estimates include as independent variables: constants, fixed effects, spatial lags of both the independent and dependent variables, and (time+space) lags of the dependent variables. They also account for spatial lags in the errors.

$***$ Significant at the 1% level.

$**$ Significant at the 5% level.

$*$ Significant at the 10% level.

Table 4

Spatial Test Statistics

(p-values in parentheses)

Statistic	(3)	(4)	(5)	(6)
	GMM initial	GMM weighted	SAR	SDM
Moran's I	0.2235 (0.00)	0.2236 (0.00)	−0.0804 (0.00)	0.7127 (0.00)
Geary	0.9054 (0.0303)	0.9055 (0.0306)	0.9946 (0.8785)	0.3561 (0.00)
Getis-Ord	−0.6392 (0.00)	−0.6394 (0.00)	0.2298 (0.00)	−2.0384 (0.00)
LM Lag (Anselin)	133.94 (0.00)	134.62 (0.00)	0.0000 (1.00)	0.1026 (0.7488)
Akaike IC	0.0122	0.0122	0.0177	0.0296

[14] See Kapoor, Kelejian and Prucha (2007).

To us, it seems clear that the SAR specification with fixed effects is preferred. It addresses the spatial autocorrelation in the errors well, though not completely. This can be seen in the reduction in Moran's I statistic (even turning it negative) and in the Lagrange multiplier test for spatial lags in the dependent variable. We conjecture that with a less crude spatial weighting matrix, the test statistics could be improved.

The conclusion that should be drawn from our empirical analysis is that some of the spatial autocorrelation in the error is due to omission of the spatially lagged dependent variable. SAR does a good (though not perfect) job of correcting this problem. The SAR model specification yields a coefficient of 0.174 on the spatially lagged dependent variable, with a standard error of 0.014.

There are two implications of this section, the first obvious and the second more subtle.

The first implication is: even if one wants to view the justification of the use of night lights data as one of a purely empirical proxy for GDP rather than a story about light demand or trade, OLS is not an appropriate specification due to spatial autocorrelation in the errors. If one inserts basic demand and trade theory into the justification, then the case for misspecification is even stronger, as there is a theoretical argument for misspecification in addition to an empirical one.

The second implication is more subtle, as it is related to how the night lights data is used in the actual calculations to proxy for GDP data. In Henderson et al (2012), the regressions of GDP on night lights given in their Table 2, namely the level regressions, are not actually used for this purpose, but rather to justify using night lights as a proxy. Until this point in this section, we have been using level regressions. Now we turn to the long difference regression, namely the regression in column (3) of Henderson et al (2012) Table (4) (reproduced in column (1) of our Table 2), which is actually used to calculate GDP proxies. In the econometric theory of these long difference regressions, what is crucial is a particular property of the OLS estimator that, when combined with sample moments, can be used to solve the parameters of the model, particularly β. That property is given in equation (9) above.[15] If our story here about model misspecification is correct, then the OLS estimator $\widehat{\psi}$ used in equation (6)

[15] As stated on p. 1007 of Henderson, Storeygard and Weil (2012), equation (9) implies that $\widehat{\psi}$ is a biased estimate of the inverse elasticity of lights with respect to income. Our claim is that even after a correction is made using equation (9) and sample moments, the estimate is still biased due to spatial autocorrelation in the errors.

should be replaced by the estimator contained in a model incorporating spatial lags in the dependent variable. Thus, the OLS estimator $\widehat{\psi}$ should be replaced by a SAR estimator. This has unknown implications for the calculations in subsequent parts of the paper that use equation (9), perhaps requiring further assumptions. With this replacement, the calculations appear to be rather intractable.

5 Implications

We have raised several issues with the analysis in Henderson et al (2012), the most important of which is the omission of electricity price, which is nowhere to be found in that paper.[16] Instead of addressing the issues one by one, we take a new approach, based on the expenditure function.

Working on empirical questions of economic growth in a national or sub-national area, is it too hard to contact a local and ask about the (average) unit price of electricity? That data can then be used in conjunction with the exchange rate and night lights data to calculate a proxy for GDP.

Why not collect data on prices and quantities of other commodities to put into the expenditure function? One could just as easily use the market for meat in addition to or in place of night lights. The expenditure function framework we have introduced explains why we might be interested in additional price and quantity data, and tells us how to incorporate it into estimates of GDP. How? Since the expenditure function is takes as domain prices and utility level, it is clear how to incorporate prices. Given a functional form, the utility level can be calculated using available quantity information, as we have done. With more data, a more flexible form of utility can be used for estimation, for example Cobb-Douglas or CES.

References

[1] Bickenbach, F., E. Bode, M. Lange and P. Nunnenkamp (2014). "Night Lights and Regional GDP," working paper http://www.ifw-members.ifw-kiel.de/publications/night-lights-and-regional-gdp.

[16] Bickenbach et al (2014) examine the stability of the coefficient of the regression of GDP on lights in the regional as opposed to national context. They find instability, which might indicate that electricity price is an omitted variable.

[2] Henderson, J.V., A. Storeygard and D.N. Weil (2012). "Measuring Economic Growth from Outer Space," *American Economic Review* 102, 994-1028.

[3] Jerison, M. (1984). "Aggregation and Pairwise Aggregation of Demand When the Distribution of Income Is Fixed," *Journal of Economic Theory* 33, 1-31.

[4] Kapoor, M., H.H. Kelejian and I. Prucha (2007). "Panel Data Models with Spatially Correlated Error Components," *Journal of Econometrics* 140, 97-130.